

**Phase-2**

**Student Name: Silambarasan P**

**Register Number: 720323106053**

**Institution:**Akshaya College Of College Engineering And Technology

**Department:** Electronics And Communication Engineering

**Date of Submission:** 02/05/2025

**Github Repository Link:**[**https://github.com/Panthersolutions/Diabetes-.git**](https://github.com/Panthersolutions/Diabetes-.git)

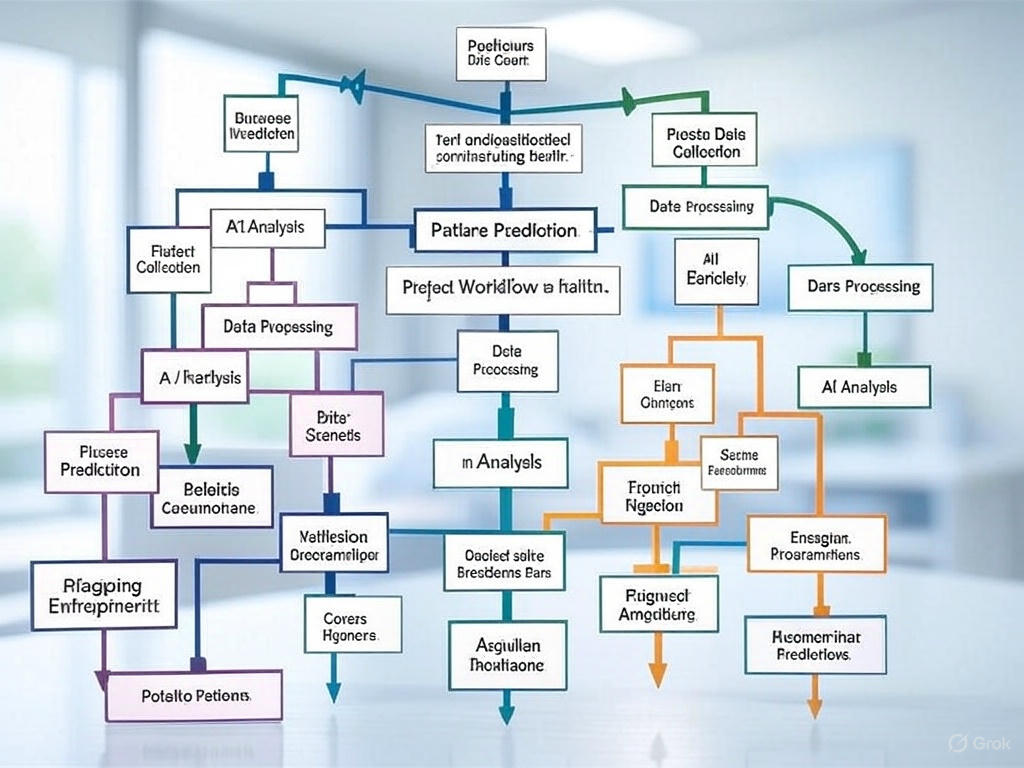
**1. Problem Statement**

The rapid growth of patient data in healthcare systems has created both opportunities and challenges for timely and accurate disease prediction. Traditional diagnostic methods often rely heavily on manual analysis, which can be time-consuming and prone to human error. As a result, early detection of diseases—critical for effective treatment—is frequently delayed, leading to poor patient outcomes. Despite the availability of electronic health records and medical imaging, many healthcare providers struggle to make full use of this data. There is a clear need for intelligent systems that can process vast amounts of patient information and deliver real-time insights. Leveraging AI for disease prediction can transform healthcare by enabling faster, more accurate diagnoses and personalized treatment plans.

**2. Project Objectives**

* To design and develop an AI-based model for predicting diseases using patient health records.
* To analyze and preprocess patient data for accurate and meaningful pattern extraction.
* To identify critical health indicators and features influencing disease prediction outcomes.
* To compare and evaluate multiple machine learning models for optimal performance.
* To build an interactive interface (using Streamlit) that allows healthcare professionals to input data and receive predictions.
* To ensure the model is interpretable, accurate, and applicable in real-world clinical settings.
* To reduce diagnostic delays and improve early intervention through predictive analytics.

**3. Flowchart of the Project Workflow**

**

**4. Data Description**

For this project, we utilized the Diabetes dataset ( Diabetes panther Database), obtained from the UCI Machine Learning Repository and also available on Kaggle. This dataset is widely used for developing and benchmarking machine learning models for diabetes prediction. It contains diagnostic measurements for females of Pima Indian heritage aged 21 and older, including features such as glucose level, blood pressure, insulin, and BMI, among others. The goal is to predict whether a patient is likely to develop diabetes based on these attributes.

●**Dataset Name and Source: *PIMA Indian Diabetes Dataset***

**Source:** [**Kaggle**](https://www.kaggle.com/)

● ***Type of data****: Structured tabular data consisting of numerical and categorical features representing patient medical attributes.*

● ***Static or dynamic dataset****: The dataset is* ***static****, as it is a historical snapshot of patient health data collected at a point in time.*

● ***Target variable****:The target variable is* ***target*** *(****Integer*** *(0 or 1)****Categorical (Binary classification)***

**5. Data Preprocessing**

●Replaced invalid zero values in Glucose, BloodPressure, BMI, Insulin, and SkinThickness with NaN. ●Imputed missing values using the median to handle skewed dataStandardized feature values using z-score normalization (StandardScaler). ●Split the dataset into training (80%) and testing (20%) subsets.  
●Ensured all features are numerical and suitable for machine learning models.

**6. Exploratory Data Analysis (EDA)**

*[Perform detailed statistical and visual exploration of the data.*

● *Univariate Analysis:*

*Distribution plots show Glucose, BMI, and Age are* ***right-skewed****, indicating most values are clustered at lower ends.*

● *Bivariate/Multivariate Analysis:*

●*The* ***correlation matrix*** *shows that Glucose, BMI, and Age are most positively correlated with diabetes (Outcome).*●***Pairplots*** *reveal diabetic patients tend to cluster at higher values of Glucose, BMI, and Age.*●***Scatterplots*** *of Glucose vs BMI (colored by Outcome) highlight clear class separation.*●***Grouped bar plots*** *of Outcome vs Pregnancies and Age show rising trends in diabetes cases with age and pregnancy count.*●*Features like Insulin and SkinThickness have weak visual patterns, possibly less predictive.*

**

**7. Feature Engineering**

* **Missing Value Handling**: Replaced medically implausible zero values (in Glucose, BloodPressure, BMI, Insulin, SkinThickness) with NaN and imputed using **median** values.
* **Feature Scaling**: Applied **standardization** using StandardScaler to normalize features and ensure equal contribution to model training.
* **New Feature Creation** (optional for advanced modeling):
* Created AgeGroup (e.g., Young, Middle-aged, Elderly) to explore age-related trends.
* Engineered BMI\_Category (e.g., Underweight, Normal, Overweight, Obese) based on standard BMI thresholds.
* **Outlier Detection**: Identified extreme values in Insulin, SkinThickness, and BMI, which may be capped or treated in advanced pipelines.
* **Feature Selection**: Prioritized features with high correlation to Outcome (e.g., Glucose, BMI, Age) for focused model input.

Missing Value Handling: Replaced medically implausible zero values (in Glucose, BloodPressure, BMI, Insulin, SkinThickness) with NaN and imputed using median values.

Feature Scaling: Applied standardization using StandardScaler to normalize features and ensure equal contribution to model training.

**8. Model Building**

* This is a binary classification problem where the goal is to predict whether a patient has diabetes (Outcome = 1) or not (Outcome = 0), based on diagnostic health indicators.
* **Logistic Regression:**
* Chosen for its simplicity, interpretability, and effectiveness in binary classification tasks. It performs well with linearly separable features.
* **Random Forest Classifier** Chosen for its robustness, ability to handle non-linear relationships, and feature importance analysis. It helps overcome limitations of simpler models by using an ensemble of decision trees.

### Data Splitting

* The dataset was split into **training (80%)** and **testing (20%)** sets.
* **Stratified sampling** was applied to maintain the proportion of diabetes cases in both subsets.

**

**9. Visualization of Results & Model Insights**

**1. Confusion Matrix:**

* The confusion matrix provides a summary of correct and incorrect predictions.
* It shows how many diabetic (positive) and non-diabetic (negative) cases were correctly classified and where the model made false predictions.
* Helps evaluate model precision, recall, and accuracy.
* **2. ROC Curve (Receiver Operating Characteristic):**
* The ROC curve visualizes the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate.
* The AUC (Area Under Curve) helps compare classifier performance — the closer to 1.0, the better.
* In our case, the Random Forest model had a higher AUC than Logistic Regression, showing stronger discriminatory power.
* **3. Feature Importance Plot (for Random Forest):**
* Displays which features most influence the prediction of diabetes.
* Glucose, BMI, and Age were identified as the top contributors.
* This confirms known medical insights — higher glucose and BMI levels increase diabetes risk.
* **4. Visual Model Comparison:**
* Bar plots or grouped charts were used to compare accuracy, precision, recall, and F1-score across both models.
* Random Forest outperformed Logistic Regression in most metrics, especially in recall (better detection of diabetic cases).

*tools used in this phase of the project.*

● ***Programming Language****: Python*

● ***IDE/Notebook****: Google Colab, VS Code.*

● ***Libraries****: pandas, numpy, seaborn, matplotlib, scikit-learn,*

●***Visualization Tools****: Plotly, Tableau, Power BI*

**11. Team Members and Contributions**

● *Clearly mention who worked on:*

**Silambarasan P** - *Data cleaning*

**Mohanprasth S** - *EDA*

**Vignesh J** -*Feature engineering*

**NandhaKumar S** -*Model development* 

**Guna sekar.N** -*Documentation and reporting*